Liu WQ, Lin MX, Huang HB *et al.* Emotion-aware music driven movie montage. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 38(3): 540-553 May 2023. DOI: 10.1007/s11390-023-3064-6

Emotion-Aware Music Driven Movie Montage

Wu-Qin Liu[1, 2] (刘伍琴), Min-Xuan Lin[3] (林敏轩), Hai-Bin Huang[3] (黄海斌), Chong-Yang Ma[3] (马重阳)
Yu Song[4] (宋　玉), Wei-Ming Dong[2, *] (董未名), *Member, CCF, ACM, IEEE*, and
Chang-Sheng Xu[2] (徐常胜), *Fellow, IEEE, Senior Member, CCF, Member, ACM*

[1] *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China*

[2] *The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

[3] *Kuaishou Technology, Beijing 100085, China*

[4] *School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China*

E-mail: liuwuqin21@mails.ucas.ac.cn; linminxuan@kuaishou.com; huanghaibin03@kuaishou.com
chongyangma@kuaishou.com; yusong@ustb.edu.cn; weiming.dong@ia.ac.cn; csxu@nlpr.ia.ac.cn

Received December 29, 2022; accepted May 22, 2023.

Abstract In this paper, we present Emotion-Aware Music Driven Movie Montage, a novel paradigm for the challenging task of generating movie montages. Specifically, given a movie and a piece of music as the guidance, our method aims to generate a montage out of the movie that is emotionally consistent with the music. Unlike previous work such as video summarization, this task requires not only video content understanding, but also emotion analysis of both the input movie and music. To this end, we propose a two-stage framework, including a learning-based module for the prediction of emotion similarity and an optimization-based module for the selection and composition of candidate movie shots. The core of our method is to align and estimate emotional similarity between music clips and movie shots in a multi-modal latent space via contrastive learning. Subsequently, the montage generation is modeled as a joint optimization of emotion similarity and additional constraints such as scene-level story completeness and shot-level rhythm synchronization. We conduct both qualitative and quantitative evaluations to demonstrate that our method can generate emotionally consistent montages and outperforms alternative baselines.

Keywords movie montage, emotion analysis, audio-visual modality, contrastive learning

1 Introduction

In recent years, with the rapid growth of social network and mobile applications, it has become increasingly popular and important to create high-quality short videos and montages. As one of the best resources for montages, movies are often cut and composed into shorter versions accompanied by a piece of background music, to obtain the trailers, previews and/or highlights of the original ones. However, existing montage editing tools typically rely on the users to manually pick shots from the movie and align with the music, which is tedious and time-consuming. It remains difficult for non-professional users to generate a movie montage of satisfactory quality to match the rhythm and emotion of the music, with the additional constraint that the selected shots provide a reasonable and comprehensible summary of the original content or story.

As machine learning technologies emerge and advance, several methods have been proposed in the past few years for the automatic generation of montages, ranging from video summarization[1] to emotion-oriented music video generation[2, 3]. However, the former mainly focuses on the content of the video itself, ignoring the correlation with any input music,

Regular Paper
Special Section of CVM 2023
This work was supported by the National Key Research and Development Program of China under Grant No. 2020AAA0106200 and the National Natural Science Foundation of China under Grant No. 61832016.
*Corresponding Author

while the latter has difficulty in understanding and handling long videos.

Walter Scott Murch, one of the most famous movie editors, has summarized the *Rule of Six* for film editing, including emotion, story, rhythm, eye trace, 2D plane of screen, and 3D space of action[4], which have different values in terms of importance for the final cut. Among these six elements, emotion is the most important one and has an importance factor of 51%, while story and rhythm correspond to a factor of 23% and 10%, respectively. Inspired by Murch's *Rule of Six*, we propose Emotion-Aware Music Driven Movie Montage (EaMD), a method to automatically generate a montage from an input movie with a piece of user-specified music as the guidance. Specifically, we compose the output montage by taking the most important three elements for film editing into account to meet the following requirements.

1) *Emotional Consistency.* The shots that are used to compose the output montage are emotionally consistent with the input music.

2) *Story Completeness.* The montage needs to present a story that is relatively complete and comprehensible.

3) *Rhythm Synchronization.* The visual and the audio content of the montage should have synchronized rhythms.

To achieve the above goals, we adopt a two-stage framework. In the first stage, we build a network to align multi-modal signals of music, text, and image in the emotion space based on CLIP (contrastive language-image pre-training) and contrastive learning[5]. In the second stage, we formulate the task of composing montages as an optimization problem and generate the output using a knapsack-based solver. Specifically, we divide the input movie at both the scene level and the shot level. The output montage is generated by maximizing the emotional similarity between scenes/shots and the input music. We ensure that the story in the montage is comprehensible by adding constraints on the number of selected scenes. Furthermore, we align selected shots with bars of the input music using quantified duration so that the rhythm of both the visual and audio signals is synchronized.

As illustrated in Fig.1, given a movie as a candidate, we can choose different shot combinations to form a montage result according to the user-supplied emotional music. The changing emotion score in the movie will be used as a significant indicator to select the target shots.

In summary, our main contributions are as follows.

• We present a novel method, EaMD, for montage generation from an input movie and a user-specified music clip based on well-established rules for film editing.

• We propose a two-stage framework to generate output movie montages, by formulating the generation task as a constrained optimization problem.

• We conduct qualitative and quantitative evaluations to demonstrate that our method leads to high-quality emotionally consistent montages and outperforms alternative baselines.

2 Related Work

Music-Driven Video Generation. The purpose of music video generation is to combine music and video to enhance entertainment quality and emotional resonance. Most previous methods, e.g., [6, 7], only con-

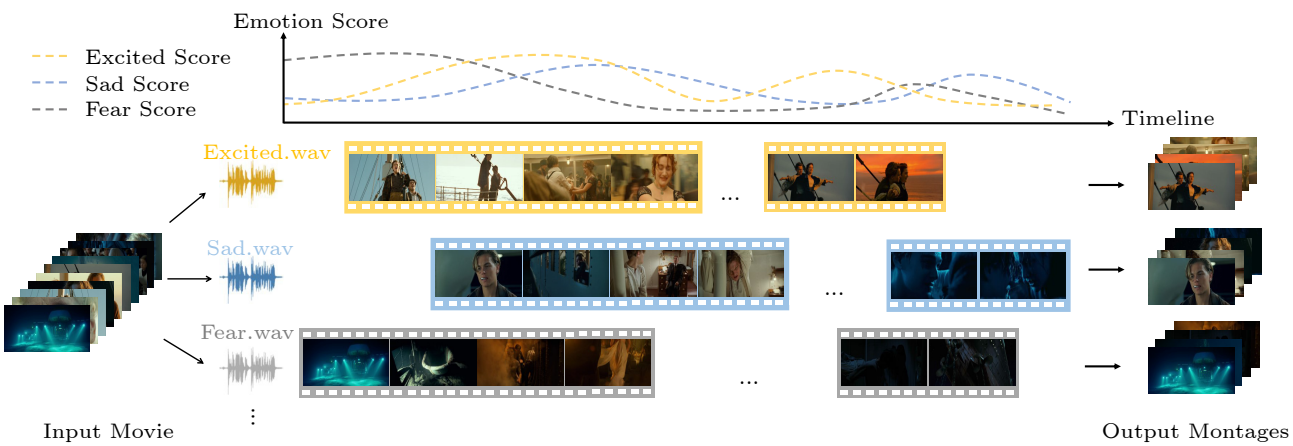


Fig.1. Emotion score based music-driven movie montage. When editing the same input movie with different background music, the corresponding emotion scores are completely different, and thus the final movie montages guided by different music are also distinct.

sider the relationship between low-level acoustic features and visual features while ignoring semantic constraints. Liao *et al.*[8] cut the input video to synchronize the music rhythm and generated audio-visually consistent results. To narrow the semantic gap between low-level acoustic features and human perception, some methods[2, 3, 9] try to map the two into the emotion space and make the audience have a good match in their emotional perception when watching the generated music video. Lin *et al.*[10] proposed an emotion-based pseudo-song prediction and matching framework. Lin *et al.*[11] considered the continuity of video content while matching music and videos. Gross *et al.*[12] generated music videos by using features of the video color histogram and key changes in music and genre. However, these methods do not consider long sequence videos. When videos are fed that are much longer than the audio time, these methods ignore the relevance of video content. They ensure emotional consistency but the generated results often lack a storyline. To address this issue, we propose an algorithm to select shots that enhance the storytelling of videos while maintaining emotional consistency.

Video Summarization. Video summarization refers to the task of generating summaries by stitching together important contents of a video. Early approaches (e.g., [13, 14]) mainly use unsupervised methods to generate video summaries due to the lack of useful datasets. After the creation of some manually collected datasets[15, 16], several supervised methods (e.g., [17]) have emerged. However, when users browse videos, they always try to find something specific. Therefore, Sharghi *et al.*[18] proposed the Query-Focused Video Summarization (QFVS) dataset, allowing video summaries to find specific shots through a query to generate results, making the results more user-friendly. After the introduction of CLIP[5], Narasimhan *et al.*[1] proposed a single framework for solving general and query-focused video summarization in both unsupervised and supervised methods by combining CLIP and video summarization. Movie trailer generation is one of the main applications of video summarization work, which attracted many researchers' attention. Existing methods usually exploit shallow audio-visual features[19–22], but these methods usually only focus on information about the movie itself. However, music is an integral part of video editing, which can affect the viewing experience of the final result. Thus, we use music as the guidance to generate an emotion-aware movie montage.

Emotion Analysis of Music and Videos. The emotions associated with music and videos have been well-studied. It has been suggested that emotions are one of the main reasons why people engage in music[23], and psychological research has shown that people also have emotional responses to visual stimuli[24]. Therefore, it is a very natural way to connect videos and music through emotions. Categorical and dimensional representations have been used to represent emotions in music[25]. Discrete categorical labels include terms like excited, relaxed, and sad. One study found that the number of emotion categories did not reflect the richness of emotions that humans perceive, or that the taxonomy is inherently ambiguous[23]. Therefore, some other studies used dimensional labels in the two-dimensional (2D) plane of valence and arousal to represent music[26]. This continuous representation has no classification problems, but it is difficult to distinguish some mental and emotional concepts. Similar to music, emotions associated with images and videos are also represented by categories[27] and dimensions[28]. Baveye *et al.*[29] expressed the features of movie scenes in the valence-arousal space. Hanjalic and Xu[30] introduced dominance as an additional dimension to characterize the emotion of videos.

3 Method

In this section, we introduce our method for emotion-aware movie montage generation. We first revisit the general setting of montage generation and then extend it into an emotion-aware constrained optimization problem. As demonstrated in Fig.2, there are two key components in our method framework, including 1) multi-modal emotion latent space alignment, and 2) emotion score based shots selection.

3.1 Problem Statement

Our goal is to generate a montage given the user-specified music x^m and a long movie x^v. Following the common practice for montage generation, we divide the movie x^v into a set of scenes $\mathcal{E} = \{e_1, e_2, ..., e_m\}$ and each scene can be split into multiple shots. We denote all the shots as a shot set $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ and use a mapping function $\tau(s_i) = e_j$ to record the relationship between scenes and shots. Similarly, the input music x^m is split into a series of bars

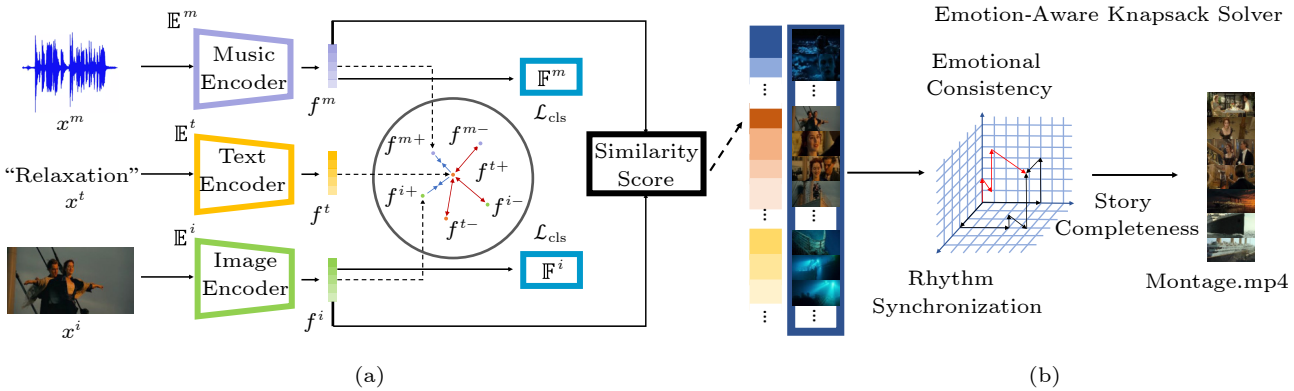


Fig.2. Illustration of our framework. (a) In the first stage, we construct an emotion space by aligning the latent representation of multiple modalities. (b) In the second stage, we select and compose several emotion-related shots from the candidates using our emotion-aware knapsack based optimization solver.

$\mathcal{B} = \{b_1, b_2, ..., b_l\}$. Then the goal of montage generation is to select a subset of shots $\mathcal{R}$ from $\mathcal{S}$ and associate each bar with a movie shot. In other words, montage generation requires 1) a shot indicator function $\mathbb{1}_s(s_i)$, determining which shots are selected and 2) a mapping function $\phi(b_k) = s_i$ to present the relationship between shots and bars. This task is in general an under-constrained problem, and hence additional constraints to $\mathbb{C} = \{c_1, c_2, ..., c_\alpha\}$ are required to limit the feasible solutions. Valid constraints include the total number of selected scenes and the rhythm synchronization between shots and bars.

In this work, we add emotion-aware constraints for the shot selection task. Our key insight is to introduce an emotion measurement function $\mathbb{M}(s_i, x^m)$, which can be used to evaluate the consistency between each shot and the whole music. With $\mathbb{M}$, we can formulate the optimization target such that the selected subset of shots $\mathcal{R}$ can construct a montage by maximizing the emotional consistency between the audio and visual signals, as shown below:

$$\mathcal{R} = \underset{i}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbb{M}(s_i, x^m)\mathbb{1}_s(s_i), \text{ s.t. } \mathbb{C}. \tag{1}$$

To solve the proposed optimization problem, we further develop a two-stage paradigm to learn the required functions. Specifically, we adopt a CLIP-based multi-modal alignment approach for emotion latent representation learning and use it as $\mathbb{M}(s_i, x^m)$. The optimization of scenes and shots selection can be modeled as a knapsack problem given the constraints $\mathbb{C}$. The shot indicator function $\mathbb{1}_s(s_i)$ and the shot-bar mapping function $\phi(b_k) = s_i$ can be obtained via a deterministic knapsack solver. We will provide details in Subsections 3.2–3.4.

3.2 Multi-Modal Emotion Latent Space Alignment

The first stage of our pipeline is to learn an emotion measurement function $\mathbb{M}$ between movie shots and music. It requires embedding and alignment of signs from different modalities in the emotion space. Inspired by AudioCLIP[31], we train three encoders ($\mathbb{E}^m$, $\mathbb{E}^t$, $\mathbb{E}^i$) of different modalities (music, text, and image) to produce matched representations, as shown in Fig.2(a). Specifically, given a tuple of music, text, and image (x^m, x^t, x^i) as the input, we use the three encoders to obtain a set of latent representations (f^m, f^t, f^i). The purpose of introducing text modality is to use it as an anchor to improve the classification accuracy. We initialize the encoders with the pretrained AudioCLIP model and further optimize joint audio-text-visual representations via contrastive learning procedure[5].

Contrastive Constraints for Multi-Modal Emotions. The pretrained AudioCLIP model gives a good embedding space for features from different modalities, and we further align the feature space and make it emotion-aware. Specifically, for arbitrary feature pair (f^a, f^b), where a, b are from different modalities, we aim to align the distribution of f^a, f^b if they correspond to the same emotion, and push away otherwise.

Towards this end, we first define an emotion indicator for different modalities. As demonstrated by Pandeya and Lee[32], emotions of audio and visual signals can be measured together in the 2D valence-arousal space, which provides a reasonable indicator to compare the differences of both modalities. Thus, we follow the settings in [32] and divide the emotion spaces into six categories to cover the emotion space

of daily communications, i.e., excited, fearful, neutral, relaxed, sad, and tense.

For each iteration of the network training stage, we build three 6×6 modality constraint matrices. Taking image-music modalities as an example, we denote a music feature and an image feature either as a positive pair (f^{m+}, f^{i+}) which will be placed on the diagonal of the matrix if they have the same emotion indicator, or as negative examples f^{i-} and f^{m-} which will be placed on the off-diagonal of the matrix. We use the cross entropy loss $\mathcal{L}_{\text{CE}}$ to push the convergence of the emotion space between two specific modalities in the matrix diagonal. The detailed image-music contrastive loss is as follows:

$$\mathcal{L}_{\text{image\_music}} = \mathcal{L}_{\text{CE}}(S^{(i+,\ m+)}, \mathbf{1}) + \mathcal{L}_{\text{CE}}(S^{(i+,\ m-)}, \mathbf{0}) + \mathcal{L}_{\text{CE}}(S^{(i-,\ m+)}, \mathbf{0}),$$

where $\mathbf{1}$ represents a full one vector, and $\mathbf{0}$ denotes a zero vector. S is the emotional consistency score we use to evaluate the distance between different modalities, defined as follows:

$$S^{(a,b)} = \frac{< f^a \times f^b >}{\|f^a\| \times \|f^b\|}.$$

We compute the constraints $\mathcal{L}_{\text{text\_image}}$ and $\mathcal{L}_{\text{text\_music}}$ for text-image modalities and text-music modalities in the same way respectively.

Emotion Classification Constraints. In order to further improve the discriminability of emotion features, we add a fully-connected layer after the image and music encoder to classify the emotion categories, which enhances the linearity of the latent emotion space. The text information is used as a tag to influence feature space construction. More concretely, a text prompt feature f^t will be reshaped to a one-hot vector $\boldsymbol{C}_{\text{text}}$ as the target. We denote the image linear classification layer as $\mathbb{F}^i$ and the music linear classification layer as $\mathbb{F}^m$. These classification layers learn to discriminate the emotion categories of images and music by the cross-entropy loss. Therefore, the final classification constraint is formulated as:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{CE}}(\mathbb{F}^i(f^i), \boldsymbol{C}_{\text{text}}) + \mathcal{L}_{\text{CE}}(\mathbb{F}^m(f^m), \boldsymbol{C}_{\text{text}}).$$

Total Loss. Our full objective loss function can be written as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{image\_music}} + \mathcal{L}_{\text{text\_music}} + \mathcal{L}_{\text{text\_image}} + \alpha \mathcal{L}_{\text{cls}},$$

where α is a parameter to balance different loss terms.

After the training stage, we apply the image encoder $\mathbb{E}^i$ on each shot to get the image feature set $F^i = \{f_1^i, f_2^i, ..., f_n^i\}$. Since each frame in a single shot is similar, we represent the content of a single shot by picking an intermediate frame x^i in the shot interval. Meanwhile, the trained music encoder $\mathbb{E}^m$ is used to extract features of the input music x^m. We collect all the emotion consistency scores for each shot and the whole music to form a set of emotion scores $\Omega = \{S_1^{(i,m)}, S_2^{(i,m)}, ..., S_n^{(i,m)}\}$ as the original value of $\mathbb{M}(s_i, x^m)$.

3.3 Emotion Score Based Shot Selection

Given the learned emotion score function $\mathbb{M}$, our next step is to select candidate shots which yield the maximum emotion score w.r.t. the optimization target in (1). Following Walter Murch's montage criterion[33], we use two constraints as $\mathbb{C}$ to limit the solution space: 1) scene-based story completeness constraint to improve the causality of the montage; 2) shot-based audio-visual rhythm synchronization constraint to guarantee the audio-visual harmonious degree of the montage. The shot indicator function $\mathbb{1}_s(s_i)$ will be obtained during optimization with these constraints.

Scene-Level Constraint for Story Completeness. Our key observation is that the less the changes in characters and the environment, the easier the audiences understand the storyline. Therefore, a high aggregation degree of scenes can provide a better story completeness. Intuitively, we can improve the story completeness by limiting the number of scenes to be involved.

Hence, we define a function $\mathbb{1}_e(e_j)$ to indicate whether a scene is selected. A scene is considered selected when one of its shots is chosen:

$$\mathbb{1}_e(e_j) = \begin{cases} 1, & \text{if } \sum\limits_{i,\tau(s_i)=e_j}^{n} \mathbb{1}_s(s_i) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, we denote $\mathbb{N}_e$ as the maximum number of selected scenes, and take it as an upper bound on the sum of $\mathbb{1}_e(e_j)$, formulated as:

$$\sum_{j=1}^{m} \mathbb{1}_e(e_j) \leqslant \mathbb{N}_e. \tag{2}$$

Shot-Level Constraint for Rhythm Synchronization. Empirically, the audiences feel more harmonious if the shot and music rhythm of a montage is changed synchronously. Here the music rhythm is defined as the duration of bars. We can model such a

rhythm synchronization constraint by establishing a mapping relationship between shots and bars. Specifically, we require each music bar should correspond to a shot and each shot should contain at least one complete music bar. To achieve this, we first quantify the duration of both shots and bars. Since the variation of music bar duration is small, we take the average continuous bar duration $\overline{t_k^{c,b}}$ as the unit of discrete time, noted as $t_k^{d,b}$. Then for each shot of movie, we obtain the discrete shot duration $t_i^{d,s}$ by exactly dividing the continuous shot duration $t_i^{c,s}$ with $t_k^{d,b}$. We further require that the sum of discrete selected shot duration is equal to the sum of all discrete bar duration $\mathbb{N}_b$, which is formulated as follows:

$$\sum_{i=1}^{n} t_i^{d,s} \mathbb{1}_s(s_i) = \sum_{k=1}^{l} t_k^{d,b} = \mathbb{N}_b. \tag{3}$$

Final Optimization Formula. We define the complete optimization problem as:

$$\begin{aligned} \mathcal{R} = \underset{i}{\operatorname{argmax}} & \sum_{i=1}^{n} \mathbb{M}(s_i, x^m) \mathbb{1}_s(s_i), \\ \text{s.t.} & \sum_{i=1}^{n} t_i^{d,s} \mathbb{1}_s(s_i) = \mathbb{N}_b, \\ & \sum_{j=1}^{m} \mathbb{1}_e(e_j) \leqslant \mathbb{N}_e. \end{aligned} \tag{4}$$

3.4 Emotion-Aware Knapsack Solver

To tackle the above optimization problem, we design an emotion-aware multi-dimensional knapsack solver with the proposed constraints. Specifically, we define three attributes belonging to shot s_i according to the optimization formula. The first one is a weighted emotion score p_i. To further enhance the importance of scenes, for each $S_i^{(i,m)}$, we adjust the value by adding the average emotion score of the scene which the shot belongs to and form the weighted emotion score set $P = \{p_1, p_2, ..., p_n\}$. Each item in P is formulated as:

$$p_i = S_i^{(i,m)} + \frac{1}{\sum_{i=1}^{n} \mathbb{1}(\tau(s_i) = e_j)} \sum_{i,\tau(s_i)=e_j}^{n} S_i^{(i,m)},$$

where $\mathbb{1}(\cdot)$ is a boolean indicator function. If the condition $(\cdot)$ holds, it returns 1, and 0 otherwise. The second attribute is the discrete shot length $t_i^{d,s}$, used to ensure the visual-audio rhythm synchronization constraint in (3). After traversing each item, we obtain the discrete shot duration set $T^{d,s} = \{t_1^{d,s}, t_2^{d,s}, ..., t_n^{d,s}\}$. The third attribute is the scene number constraint score q_i corresponding to (2). We define a step function relying on the subscripts of the scene to which the shot belongs, used to classify different scene categories:

$$q_i = j, \ e_j = \tau(s_i).$$

We denote $Q = \{q_1, q_2, ..., q_n\}$ as the set of scores for scene number constraints. The three attribute sets will be regarded as individual factors in the knapsack solver.

Hard Scene Constraint Knapsack Solver. The basic state (i, j, k, z) is defined to represent the maximum emotion score achieved by selecting exactly k scenes, with a total discrete shot duration of j, and iterating over the first i shots. z means whether the scene to which the i-th shot belongs is selected. We display the detailed state transition equation in Algorithm 1. Considering whether the current state is on the boundary of the scene $(q_i \neq q_{i-1})$, four possible state transition paths need to be discussed separately. When the user queries a specific upper bound on the number of scenarios $\mathbb{N}_e$, the maximum emotion score $\mathbb{P}$ can be quickly looked up. Meanwhile, the $Backtrack(\cdot)$ method, as the shot indicator function $\mathbb{1}_s(s_i)$, will trace a legal path in inverse order and return a possible index set of shots R. Then, we can obtain the mapping function $\phi(b_k)$ by matching selected shots and bars of music in chronological order.

Algorithm 1. Hard Scene Constraint Knapsack Solver

Input: set P, set Q, set $T^{d,s}$, and $n, \mathbb{N}_b, \mathbb{N}_e$ as corresponding capacity

Output: the maximum emotion score $\mathbb{P}$, the picked shot index set R

for $i : 1 \to n$ **do**
  for $j : 1 \to \mathbb{N}_b$ **do**
    for $k : 1 \to \mathbb{N}_e$ **do**
      if $q_i \neq q_{i-1}$ **then**
        $(i, j, k, 1) \leftarrow \max((i-1, j - t_i^{d,s}, k-1, 0) + p_i, (i-1, j - t_i^{d,s}, k-1, 1) + p_i)$
        $(i, j, k, 0) \leftarrow \max((i-1, j, k, 0), (i-1, j, k, 1))$
      else
        $(i, j, k, 1) \leftarrow \max((i-1, j - t_i^{d,s}, k-1, 0) + p_i, (i-1, j - t_i^{d,s}, k, 1) + p_i, (i-1, j, k, 1))$
        $(i, j, k, 0) \leftarrow (i-1, j, k, 0)$
      end if
    end for
  end for
end for
$\mathbb{P} \leftarrow \max((n, \mathbb{N}_b, \mathbb{N}_e, 1), (n, \mathbb{N}_b, \mathbb{N}_e, 0))$
$R \leftarrow Backtrack(\mathbb{P})$
return $\mathbb{P}, R$

The hard scene constraint requires that the total number of selected scenes is less than an upper bound. As illustrated in Fig.3(a), we constrain the capacity of the set of scenes to which selected shots belong. Algorithm 1 in the main paper displays the details of hard scene constraint knapsack. We iterate through all possible states with a triple loop which contains three core factors. In each state transition, the current state will obtain the maximum emotion score from some legal substates. Specifically, four different state transition cases need to be discussed.

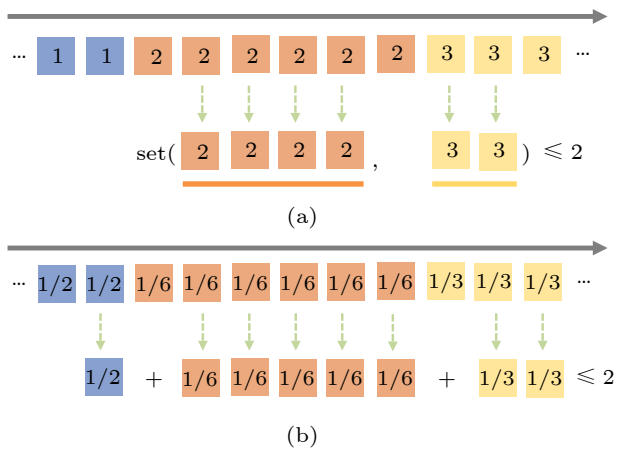


Fig.3. (a) Hard scene constraints. The sum of weights of all picked shots is limited. (b) Soft scene constraints. The number of scenes that picked shots belong to is constrained.

1) If q_i and q_{i-1} belong to different scenes, we choose the scene to which the i-th shot belongs ($z = 1$). In line 5 of Algorithm 1, two valid substates that reduce the number of scenes should be considered when selecting a scene on the boundary.

2) If q_i and q_{i-1} belong to the different scenes, we do not choose the scene to which the i-th shot belongs ($z = 0$). In line 6 of Algorithm 1, the scene number will not decrease in substates because nothing is selected.

3) If q_i and q_{i-1} belong to the same scene, we choose the scene to which the i-th shot belongs ($z = 1$). Line 8 of Algorithm 1 shows three possible substates. If the i-th shot is selected, the algorithm needs to separately consider whether the number of scenes is reduced. Conversely, the scene to which the i-th shot belongs must be chosen before this state.

4) If q_i and q_{i-1} belong to the same scene, we do not choose the scene to which the i-th shot belongs ($z = 0$). In line 9 of Algorithm 1, the scene to which the i-th shot belongs cannot be selected in the substate.

Soft Scene Constraint Knapsack Solver. As illustrated in Fig.3(b), the soft scene constraint knapsack solver assigns corresponding scene constraint weight for each shot and limits the sum of weights for all selected shots. Before starting optimization, we multiply all scene constraint weights by a magnification constant and round them down to ensure each weight is an integer.

Fixing the number of scenes may fail to obtain the highest sum of emotion scores. Thus, we loosen the restriction in (2). Instead of limiting the upper bound of the sum of selected scenes, we constrain that the sum of the inverse of the number of shots in the scene to which the selected shots belong is no larger than $\mathbb{N}_e$:

$$\sum_{i=1}^{n} \frac{1}{\sum_{\alpha=1}^{n} \mathbb{1}(\tau(s_\alpha) = \tau(s_i))} \mathbb{1}_s(s_i) \leqslant \mathbb{N}_e,$$

where $\mathbb{1}(\cdot)$ is a standard indicator. If the equation is established, the function value is 1; otherwise it is 0.

Then, we reconstruct the soft scene number constraint score set as $\tilde{Q} = \{\tilde{q_1}, \tilde{q_2}, ..., \tilde{q_n}\}$, where each item of the set is formulated as:

$$q_i = \frac{1}{\sum_{\alpha=1}^{n} \mathbb{1}(\tau(s_\alpha) = e_j)}, \ e_j = \tau(s_i).$$

In this condition, Algorithm 1 will degenerate into a vanilla three-dimensional knapsack solver. We assume a basic state (i, j, k), which stores the maximum emotion score when it traverses to the i-th shot constrained by the sum of picked scene weight j and the sum of picked discrete shot duration k. During optimization, the state (i, j, k) will visit all $(i-1, j-t_i^{d,s}, k-q_i)$ states, and pick the maximum value to transfer. We get the same results as above.

At last, when the function of searching the best solution $\mathbb{1}_s(s_i)$ has been obtained, we discard the part where the shot is longer than the bar to align duration, concatenate all selected shots in chronological order and append the given music according to $\phi(b_k)$ to get the final montage. In general, we provide two knapsack-based deterministic optimization schemes to select the shot with high emotion relevance from abundant candidate shots.

4 Experiments

4.1 Dataset

The music video dataset[32] is used to train our

model. This dataset focuses on the multimodal emotion classification task, utilizing audio and visual information to discriminate the category of music videos. During the training stage, a total of 4 788 samples are used, consisting of videos that convey 843 instances of excitement, 828 instances of fear, 678 instances of neutrality, 1 057 instances of relaxation, 730 instances of sadness, and 652 instances of tension. Since the labels are assigned to the whole video, we assume that the emotion of each frame within the same shot is consistent in each batch of training. We randomly pick a frame from videos as the input of image encoder, and use the full music to encode the audio feature. For text modality, we use six fixed text prompts. Finally, we test the generated results on a test set of 300 samples, where each emotion category contains 50 videos.

The original data in the dataset we use has consistent and rich emotions. Concretely, the consistency represents the raw materials convey the same emotion signal in the visual and audio modalities. For example, the "excited" contains positive emotions with bright hued scenes and the corresponding music has light rhythm and pleasant chords. Meanwhile, the richness of emotions means that each category in the dataset covers various fine-grained emotions. For example, the category of "excited" encompasses emotions such as happiness, joy, love, and excitement, while the category of "fear" includes emotions such as fearfulness, disgust, terror, and so on.

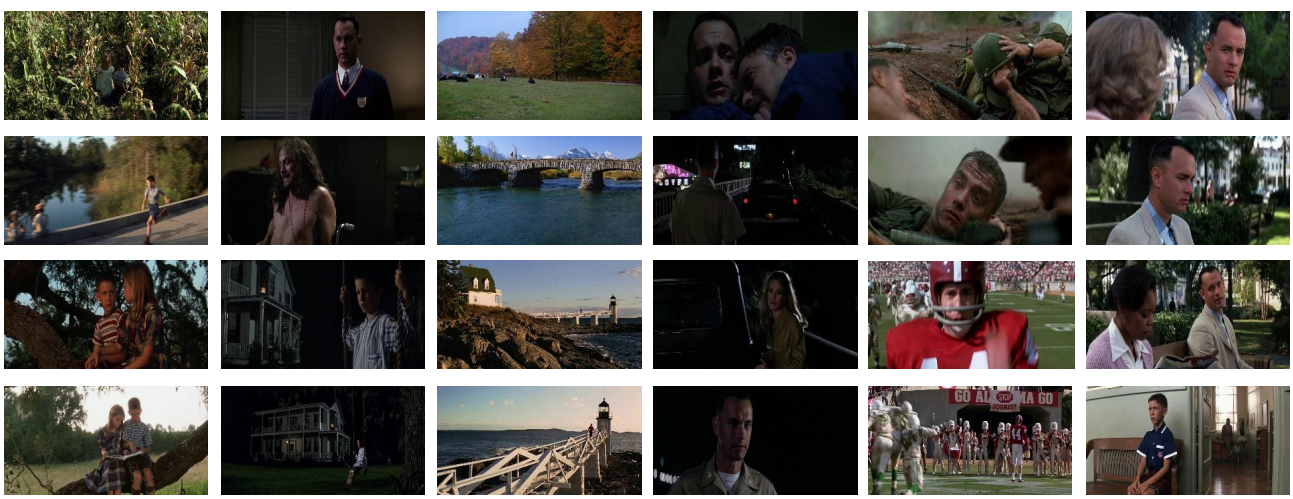
4.2 Experimental Setup

For visual modality, we extract the shot of video by TransNet v2[34] and obtain the scene segmentation boundary by the method of Rao *et al.*[35]. For audio modality, we split the bar of music by the Madmom library[36]. We train our model for 50 epochs with the Adam optimizer[37] on a single NVIDIA RTX 3090. The learning rate is 0.000 1 and the batch size is 6. Meanwhile, we set the trade-off α as 1. In the optimization stage, we denote Ours(h) as the montage results generated by using the hard scene constraint, and Ours(s) as the ones generated by using the soft constraint. The scene number constraint for both methods is 5.

To comprehensively compare the differences between various types of movies and music in the montage task, in the evaluation phrase, we choose 11 movies whose categories cover action (e.g., L*é*on), love (e.g., Titanic), science fiction (e.g., Inception), comedy (e.g., The Grand Budapest Hotel) and fear (e.g., Train to Busan). Sixteen pieces of music with distinct emotions are used as background songs.

4.3 Qualitative Evaluations

In this subsection, we show some qualitative results from visual-audio aspect. To directly evaluate the quality of montage, Fig.4 shows some visualized results that frames are picked from the montage generated by our method. Forrest Gum, as an example, is clipped by various kinds of music with different emotions. The representative pictures with strong emotions are shown in each column. Apparently, some optimistic scenes (excitement or relaxation)



(a) (b) (c) (d) (e) (f)

Fig.4. Montage results generated by our framework driven by music with different emotions. (a) Excited.wav. (b) Fear.wav. (c) Relaxation.wav. (d) Sad.wav. (e) Tension.wav. (f) Neutral.wav.

with the bright light are selected by the delighted music. On the other side, the painful scenes (fear or sadness) are often accompanied by crying and dark atmosphere. To some extent, the results demonstrate that our method has a good audio-visual emotional consistency performance.

Ablation Study. To explore the impact of each component in our solvers on the preferences of the audience, we ablate three key factors in (4), including emotional consistency, story completeness and rhythm synchronization, to make a 30-second montage with fixed music. For user study, we select five movies and generate five montages for each movie by our models and the baseline (ablated) models. We also invite an expert to make a montage for each movie under same conditions. Finally, we get 30 montages and invite 36 investigators to rate them (1–5), considering four aspects: 1) the degree of audio-visual emotional consistency; 2) the degree of story completeness; 3) the degree of audio-visual rhythm synchronization; 4) the overall quality of the montage.

Table 1 shows the average rating statistics. Apart from the results from the expert as upper bound, Ours(h) achieves the highest rating in story completeness, rhythm synchronization and overall evaluation under full constraints. We bold the highest score of our method in each metric. With a movie of about two hours as a benchmark, our method only takes 20 minutes to process a montage, but it takes the expert 2–3 days to process a 30-second video, because it takes a lot of time to choose suitable shots. Further, by relaxing the constraint of the scenes, Ours(s) outperforms on emotional consistency than Ours(h) but slightly decreases in other metrics due to the loss of overall coherence. When the emotion factor is not considered (w/o emotion), there is a significant drop in all ratings, proving the importance of audio-visual emotional consistency for montages. Similarity, despite the selection of the largest emotion score, the lack of a story completeness constraint (w/o story) will limit the overall quality of montages. Due to people's sensitivity to audio-visual rhythm synchronization (sync.), the last factor (w/o rhythm) gets almost the worst score in most aspects.

Table 1. Results of Ablation Study

Method	Emotional Consistency	Story Completeness	Rhythm Sync.	Overall
Ours(h)	3.574	**3.624**	**3.616**	**3.783**
Ours(s)	**3.672**	3.148	3.438	3.502
w/o emotion	3.026	3.146	3.412	3.105
w/o story	3.384	3.140	3.460	3.328
w/o rhythm	3.182	3.044	3.124	3.138
Expert	3.938	3.886	3.966	4.037

Qualitative Comparisons with Other Methods. To the best of our knowledge, the proposed framework is the first to achieve music-driven movie montage, lacking comparable methods and open source codes. [11] by Lin *et al.* is the most similar work to ours, which firstly recommends a piece of matched music from a fixed music database according to the user-supplied video and then obtains the final montage by selecting shots under cost-based constraints. Although the input is not exactly consistent, the output of that work[11] is the same as ours; therefore, we set it as a baseline. To further prove the effectiveness of our approach, we invite the expert to clip the montages under the same conditions.

In this study, we make montages of lengths between three and five minutes with a piece of music from the user-specified video about 15 minutes. Finally, we produce five montages and invite 40 investigators to participate this experiment. These participants receive the same questions as the user study. They rate each montage on a scale of 1–5 based on emotional consistency, story completeness and rhythm synchronization and overall aspect. Table 2 demonstrates that compared with [11], we achieve significant superiority on all evaluation metrics. We bold the highest score of methods in each evaluation criterion except for expert ratings. In particular, since we explicitly consider the influencing factors of film editing, we achieve large improvements in story completeness and rhythm synchronization, and can even achieve scores that are competitive with expert results.

Table 2. Qualitative Comparisons with Other Methods

Method	Emotional Consistency	Story Completeness	Rhythm Sync.	Overall
Lin *et al.*[11]	3.690	3.723	3.178	3.401
Ours	**3.782**	**4.101**	**3.678**	**3.987**
Expert	3.835	3.948	4.024	4.103

Arbitrary Music Driven Movie Montage. To reflect the disparity guided by different types of music, we conduct a user study to explore the audio-visual emotional consistency of our approach. For comparison, we select different types of pieces of emotional music to drive movie clips, and each movie will be

edited by two random emotions. In this experiment, investigators need to answer two questions. 1) What emotions do you feel from the movie montage? 2) How strong are they? Considering different proportions of emotions of diverse people, we allow them to choose multiple emotions for each movie montage and score the emotion degree in the montage on a scale of 1–10, where higher scores represent more prominent emotions. Finally, we receive a total of 46 valid questionnaires. As shown in Fig.5, in each row, we display the voted percentage of each emotion category for a single movie driven by a piece of music with different emotions. For each piece of music, we draw the normalized degree of relevance that the participants voted on, and the red word means the highest probability category. By observing the results, we achieve the distinct differences in all six emotions. The relaxation is the easiest category to tell due to that beautiful landscapes and bright scenes are often appeared. The excitement and tension become the most confusing emotion category on account of a large amount of similar facial expressions and body movements.

4.4 Quantitative Evaluations

Confusion Matrix of Emotion Classification. We apply the music video emotion classification accuracy to assess the performance of our model. We validate our model on the test set using a confusion matrix as a visual evaluation method, which counts the number of samples in classes that are misclassified between different categories. As shown in Fig.6, our model performs well on categories of "fear", "relaxation" and "excited". However, "neutrality" is often misclassified as other classes, because its data resembles other emotions.

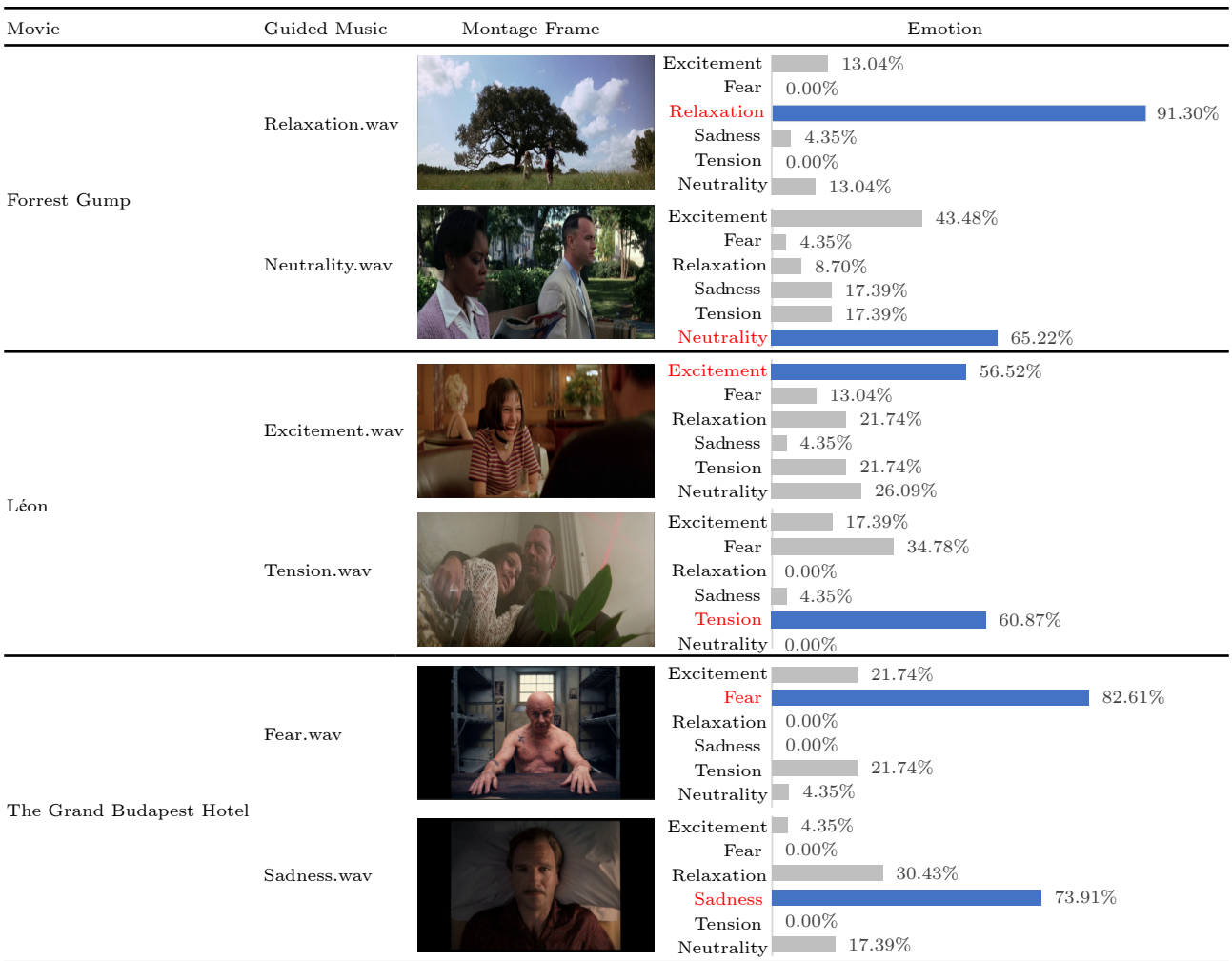


Fig.5. Percentage of most possible emotions for movie montages guided by two pieces of random emotional music. Six examples are displayed, including the guided music, the montage frame, and the emotion probability map voted by investigators.

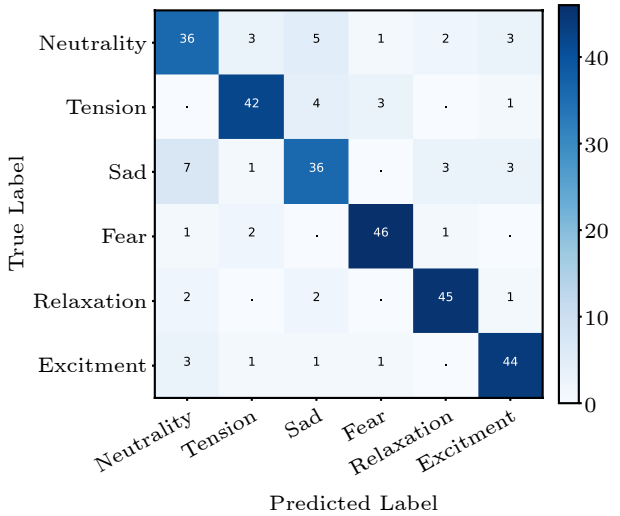


Fig.6. Confusion matrix of emotion classification.

Statistics of the Accuracy and F1-Score. We compare our methods with others in terms of top-1 accuracy (Acc.) and $F1$-score (F_1) to prove the discriminability of our emotion space by feeding signals in different modalities. We select the method of [38], vanilla AudioCLIP[31], and Wav2CLIP[39] as baselines. The results are shown in Table 3, where "Ours w cls" and "Ours w/o cls" correspond to training models with and without a classifier, respectively. "Ours w/o text-enc" means to remove the text encoder and does not use the text modality to enhance the feature. "Ours w/o pretrain" means the encoder is trained from scratch. We demonstrate the effectiveness of our full framework by comparing the classification performance of encoders under various conditions. We bold the highest score and underline the second highest score in each metric. We achieve the best top-1 accuracy and $F1$-score on the emotion classification task of music videos. AudioCLIP and Wav2CLIP completely lose the ability to classify video emotions due to the constraint of the original pretrained dataset. Compared with [38], our method also achieves the highest performance in the audio modality.

Table 3. Statistics of the Accuracy and $F1$-Score

Method	Accuracy (%)		F_1 (%)	
	Audio	Visual	Audio	Visual
Pandeya *et al.*[38]	74.0	**74.0**	73.0	**73.0**
AudioCLIP[31]	18.3	34.0	12.6	32.9
Wav2CLIP[39]	16.3	12.7	4.7	11.5
Ours w cls	**82.0**	<u>69.7</u>	**82.0**	<u>70.0</u>
Ours w/o cls	<u>76.7</u>	67.7	<u>76.9</u>	67.7
Ours w/o text-enc	75.3	63.2	75.1	63.2
Ours w/o pretrain	69.0	56.8	69.2	56.6

4.5 Video Demo

To demonstrate the effectiveness of our framework, we provide a video demo that consists of movie montages guided by various pieces of emotional music. Two specific tasks are presented in the video.

The first task is to generate movie montages driven by different pieces of emotional music for a single movie. We list two different movies, including "Forrest Gump" and "Leon". For the same movie, we process it with our pipeline, adding a piece of 30-second emotional music, such as "Forrest Gump" edited by relaxed and neutral music and "Leon" with excited and tense music. Apparently, we can easily observe the difference in the montage results. For example, the beautiful landscape (i.e., sea and forest) frames are mainly picked in the montage when a music with the relaxed emotion is used as guidance. On the contrary, with a piece of neutral music, the movie montage often contains static pictures, for example, the expressionless man sitting on the chair. Based on our solver, we successfully select the suitable set of movie shots to create a montage, which leads to the disparity.

The second task in our demo is to create the montage using the corresponding theme song of the movie. Two movies, "Mulan" and "The Grand Budapest Hotel", are used as raw materials. In this case, we show our framework can create a montage that fits the overall mood and rhythm of the movie according to the theme song.

5 Discussions and Limitations

Although our method can achieve emotional consistency and rhythmic synchronization in montage, there are still some limitations. For example, when the selected shot length is longer than the corresponding music bar length, part of the shot will be cut off, which may compromise the integrity of the shot. Moreover, our method does not explicitly consider the semantic transition and coherence between shots; thus there may be plot jumps between two consecutive shots. In the future, we will further explore and address these issues.

6 Conclusions

In this paper, we introduced a new task for emotional perception in movie montages that is driven by music. Our task involves the challenge of retrieving

and reorganizing shots from movies based on specific segments of music selected by the user. To address this problem, we formulated it as an optimization problem and proposed a two-stage framework that includes a learning-based module for predicting emotional similarity and an optimization-based module for selecting and synthesizing candidate movie shots.

Our proposed framework is effective in capturing movie shots that align with audio emotions and generating storytelling montage videos. We conducted qualitative and quantitative evaluations, and achieved the highest accuracy of 82% and $F1$-score of 82%, which demonstrate that our method can handle ultra-long videos and produce high-quality results compared with existing methods.

In the future, we plan to further improve our approach in both video transitions and storytelling to achieve results comparable to those of professional editors.

References

[1] Narasimhan M, Rohrbach A, Darrell T. Clip-it! Language-guided video summarization. In *Proc. the 35th International Conference on Neural Information Processing Systems*, Dec. 2021, pp.13988–14000.

[2] Lin J C, Wei W L, Wang H M. EMV-matchmaker: Emotional temporal course modeling and matching for automatic music video generation. In *Proc. the 23rd ACM International Conference on Multimedia*, Oct. 2015, pp.899–902. DOI: 10.1145/2733373.2806359.

[3] Lin J C, Wei W L, Wang H M. DEMV-matchmaker: Emotional temporal course representation and deep similarity matching for automatic music video generation. In *Proc. the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2016, pp.2772–2776. DOI: 10.1109/ICASSP.2016.7472182.

[4] Murch W. In the Blink of an Eye. Silman-James Press, 2001.

[5] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In *Proc. the 38th International Conference on Machine Learning*, Jul. 2021, pp.8748–8763.

[6] Yoon J C, Lee I K, Byun S. Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools and Applications*, 2009, 41(2): 197–214. DOI: 10.1007/s11042-008-0225-0.

[7] Kuo F F, Shan M K, Lee S Y. Background music recommendation for video based on multimodal latent semantic analysis. In *Proc. the 2013 IEEE International Conference on Multimedia and Expo*, Jul. 2013. DOI: 10.1109/ICME.2013.6607444.

[8] Liao Z C, Yu Y Z, Gong B C, Cheng L C. Audeosynth: Music-driven video montage. *ACM Trans. Graphics*, 2015, 34(4): Article No. 68. DOI: 10.1145/2766966.

[9] Wang J C, Yang Y H, Jhuo I H, Lin Y Y, Wang H M. The acousticvisual emotion Guassians model for automatic generation of music video. In *Proc. the 20th ACM International Conference on Multimedia*, Oct. 2012, pp.1379–1380. DOI: 10.1145/2393347.2396494.

[10] Lin J C, Wei W L, Wang H M. Automatic music video generation based on emotion-oriented pseudo song prediction and matching. In *Proc. the 24th ACM International Conference on Multimedia*, Oct. 2016, pp.372–376. DOI: 10.1145/2964284.2967245.

[11] Lin J C, Wei W L, Yang J, Wang H M, Liao H Y M. Automatic music video generation based on simultaneous soundtrack recommendation and video editing. In *Proc. the 25th ACM International Conference on Multimedia*, Oct. 2017, pp.519–527. DOI: 10.1145/3123266.3123399.

[12] Gross S, Wei X X, Zhu J. Automatic realistic music video generation from segments of YouTube videos. arXiv: 1905.12245, 2019. https://arxiv.org/abs/1905.12245, Jun. 2023.

[13] Liu T C, Kender J R. Optimization algorithms for the selection of key frame sequences of variable length. In *Proc. the 7th European Conference on Computer Vision*, Apr. 2002, pp.403–417. DOI: 10.1007/3-540-47979-1_27.

[14] Mahmoud K M, Ghanem N M, Ismail M A. Unsupervised video summarization via dynamic modeling-based hierarchical clustering. In *Proc. the 12th International Conference on Machine Learning and Applications*, Dec. 2013, pp.303–308. DOI: 10.1109/ICMLA.2013.140.

[15] Song Y L, Vallmitjana J, Stent A, Jaimes A. TVSum: Summarizing web videos using titles. In *Proc. the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.5179–5187. DOI: 10.1109/CVPR.2015.7299154.

[16] Gygli M, Grabner H, Riemenschneider H, Van Gool L. Creating summaries from user videos. In *Proc. the 13th European Conference on Computer Vision*, Sept. 2014, pp.505–520. DOI: 10.1007/978-3-319-10584-0_33.

[17] Zhu W C, Lu J W, Li J H, Zhou J. DSNet: A flexible detect-to-summarize network for video summarization. *IEEE Trans. Image Processing*, 2021, 30: 948–962. DOI: 10.1109/TIP.2020.3039886.

[18] Sharghi A, Gong B Q, Shah M. Query-focused extractive video summarization. In *Proc. the 14th European Conference on Computer Vision*, Oct. 2016, pp.3–19. DOI: 10.1007/978-3-319-46484-8_1.

[19] Irie G, Satou T, Kojima A, Yamasaki T, Aizawa K. Automatic trailer generation. In *Proc. the 18th ACM International Conference on Multimedia*, Oct. 2010, pp.839–842. DOI: 10.1145/1873951.1874092.

[20] Xu H T, Zhen Y, Zha H Y. Trailer generation via a point process-based visual attractiveness model. In *Proc. the 24th International Conference on Artificial Intelligence*, Jul. 2015, pp.2198–2204. DOI: 10.5555/2832415.2832554.

[21] Papalampidi P, Keller F, Lapata M. Film trailer genera-

tion via task decomposition. arXiv: 2111.08774, 2021. https://arxiv.org/abs/2111.08774, Jun. 2023.

[22] Smith J R, Joshi D, Huet B, Hsu W, Cota J. Harnessing A I. for augmenting creativity: Application to movie trailer creation. In *Proc. the 25th ACM International Conference on Multimedia*, Oct. 2017, pp.1799–1808. DOI: 10.1145/3123266.3127906.

[23] Juslin P N, Laukka P. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 2004, 33(3): 217-238. DOI: 10.1080/09298210420 00317813.

[24] Cowen A S, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, 2017, 114(38): E7900-E7909. DOI: 10.1073/pnas.1702247114.

[25] Kim Y E, Schmidt E M, Migneco R, Morton B G, Richardson P, Scott J J, Speck J A, Turnbull D. Music emotion recognition: A state of the art review. In *Proc. the 11th International Society for Music Information Retrieval Conference*, Aug. 2010, pp.937–952.

[26] Russell J A. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980, 39(6): 1161–1178. DOI: 10.1037/h0077714.

[27] Zhao S C, Gao Y, Jiang X L, Yao H X, Chua T S, Sun X S. Exploring principles-of-art features for image emotion recognition. In *Proc. the 22nd ACM International Conference on Multimedia*, Nov. 2014, pp.47–56. DOI: 10.1145/2647868.2654930.

[28] Lu X, Suryanarayan P, Adams Jr R B, Li J, Newman M G, Wang J Z. On shape and the computability of emotions. In *Proc. the 20th ACM International Conference on Multimedia*, Oct. 2012, pp.229–238. DOI: 10.1145/2393347.2393384.

[29] Baveye Y, Dellandrea E, Chamaret C, Chen L M. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Trans. Affective Computing*, 2015, 6(1): 43–55. DOI: 10.1109/TAFFC.2015.2396531.

[30] Hanjalic A, Xu L Q. Affective video content representation and modeling. *IEEE Trans. Multimedia*, 2005, 7(1): 143–154. DOI: 10.1109/TMM.2004.840618.

[31] Guzhov A, Raue F, Hees J, Dengel A. Audioclip: Extending clip to image, text and audio. In *Proc. the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2022, pp.976–980. DOI: 10.1109/ICASSP43922.2022.9747631.

[32] Pandeya Y R, Lee J. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 2021, 80(2): 2887–2905. DOI: 10.1007/s11042-020-08836-3.

[33] Yu X J. A study on the editing frequencies trends for films emotion clips. *International Journal of Organizational Innovation*, 2017, 9(3): 40A–47A.

[34] Souček T, Lokoč J. TransNet V2: An effective deep network architecture for fast shot transition detection. arXiv: 2008.04838, 2020. https://arxiv.org/abs/2008.04838, Jun. 2023.

[35] Rao A Y, Xu L N, Xiong Y, Xu G D, Huang Q Q, Zhou B L, Lin D H. A local-to-global approach to multi-modal movie scene segmentation. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.10146–10155. DOI: 10.1109/CVPR42600.2020.01016.

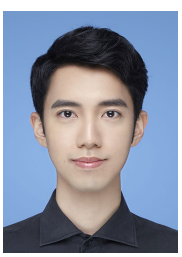
[36] Böck S, Korzeniowski F, Schlüter J, Krebs F, Widmer G. Madmom: A new python audio and music signal processing library. In *Proc. the 24th ACM International Conference on Multimedia*, Oct. 2016, pp.1174–1178. DOI: 10.1145/2964284.2973795..

[37] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv: 1412.6980, 2014. https://arxiv.org/abs/1412.6980, Jun. 2023.

[38] Pandeya Y R, Bhattarai B, Lee J. Deep-learning-based multimodal emotion classification for music videos. *Sensors*, 2021, 21(14): 4927. DOI: 10.3390/s21144927.

[39] Wu H H, Seetharaman P, Kumar K, Bello J P. Wav2CLIP: Learning robust audio representations from clip. In *Proc. the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2022, pp.4563–4567. DOI: 10.1109/ICASSP43922.2022.9747669.

Wu-Qin Liu received her B.S. degree in communication engineering from Beijing Jiaotong University, Beijing, in 2021. She is currently a graduate student at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing. Her research interest is multimodal visual media generation.



Min-Xuan Lin received his M.S. degree in computer applied technology from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, in 2021. He received his B.S. degree in computer science and technology from Ocean University of China, Qingdao, in 2018. He is currently an engineer at Kuaishou Technology, Beijing. His research interests include computational visual media and visual media generation.

Hai-Bin Huang received his B.S. and M.S. degrees in mathematics in 2009 and 2011 respectively from Zhejiang University, Hangzhou. He obtained his Ph.D. degree in computer science from UMass Amherst, Amherst, in 2017. He is currently a senior staff research scientist at Kuaishou Technology, Beijing. His research interests include computer graphics.

Chong-Yang Ma received his B.S. degree from the Fundamental Science Class (Mathematics and Physics) of Tsinghua University, Beijing, in 2007, and his Ph.D. degree in computer science from the Institute for Advanced Study of Tsinghua University, Beijing, in 2012. He is currently a research manager at Kuaishou Technology, Beijing. His research interests include computer graphics and computer vision.

Yu Song received her Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, and her M.S. degree in automation from Tianjin University, Tianjin, in 2017. She is currently working in the Industrial Design Department of the School of Mechanical Engineering, University of Science and Technology Beijing, Beijing. Her research interests include image collage, image retargeting, and machine learning

Wei-Ming Dong is a professor at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS) at Institute of Automation, Chinese Academy of Sciences, Beijing. He received his B.S. and M.S. degrees in computer science and technology in 2001 and 2004, respectively, both from Tsinghua University, Beijing. He received his Ph.D. degree in computer science from the University of Lorraine, Nancy, in 2007. His research interests include image synthesis, image recognition, and computational creativity.

Chang-Sheng Xu is a professor at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, and the executive director of the China-Singapore Institute of Digital Media, Singapore. Prof. Xu received the Best Associate Editor Award of ACM Transactions on Multimedia Computing, Communications and Applications in 2012 and the Best Editorial Member Award of ACM/Springer Multimedia Systems Journal in 2008. He served as a program chair of the Association for Computing Machinery (ACM) Multimedia 2009. He has served as an associate editor, a guest editor, a general chair, a program chair, an area/track chair, a special session organizer, a session chair, and a Transactions on Professional Communication (TPC) member for over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is an associate editor of ACM Transactions on Multimedia Computing, Communications and Applications and the editor-in-chief of ACM/Springer Multimedia Systems Journal. He is a fellow of IEEE and IAPR, a senior member of CCF, and a member of ACM.